# Knowledge Discovery by Relation Approximation: A Rough Set Approach

Hung Son Nguyen

Institute of Mathematics
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
son@mimuw.edu.pl

## Extended Abstract

In recent years, rough set theory [1] has attracted attention of many researchers and practitioners all over the world, who have contributed essentially to its development and applications. With many practical and interesting applications rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition [2].

The common issue of the above mentioned domains is the concept approximation problem which is based on searching for description – in a predefined language $\mathcal{L}$ – of concepts definable in other language $\mathcal{L}^*$. Not every concept in $\mathcal{L}^*$ can be exactly described in $\mathcal{L}$, therefore the problem is to find an approximate description rather than exact description of unknown concepts, and the approximation is required to be as exact as possible. Usually, concepts are interpretable as subsets of objects from a universe, and the accuracy of approximation is measured by the closeness of the corresponding subsets.

Rough set theory has been introduced as a tool for concept approximation from incomplete information or imperfect data. The essential idea of rough set approach is to search for two descriptive sets called *the lower approximation* containing those objects that certainly belong to the concept and the "upper approximation" containing those objects that possibly belong to the concept.

Most concept approximation methods realize the inductive learning approach, which assumes that a partial information about the concept is given by a finite sample, so called *the training sample or training set*, consisting of positive and negative cases (i.e., objects belonging or not belonging to the concept). The information from training tables makes the search for patterns describing the given concept possible. In practice, we assume that all objects from the universe $\mathcal{U}$ are perceived by means of information vectors being vectors of attribute values (information signature). In this case, the language $\mathcal{L}$ consists of boolean formulas defined over conditional (effectively measurable) attributes.

The task of concept approximation is possible when some information about the concept is available. Except the partial information above the membership function given by training data set, the domain knowledge is also very useful

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 103–106, 2006.

in developing efficient methods of searching for accurate approximate models. Unfortunately, there are two major problems related to the representation and the usage of the domain knowledge can cause many troubles in practical applications. In [3] [4] [5] we have presented a method of using domain knowledge, which is represented in form of concept taxonomy, to improve the accuracy of rough classifiers and to manage with approximation problems over complex concepts. The proposed solution adopts the general idea of multi-layered learning approach [6] where the original problem is decomposed into simpler ones and the main effort is to synthesize the final solution from the solutions of those simpler problems.

Usually rough set methodology is restricted to decision tables and is destined to classification task. This paper focus on applications of rough sets and layered learning in other KDD tasks like approximation of concept defined by decision attribute with continuous domain or ranking learning.

In mathematics, $k$-argument relations over objects from a given universe $\mathcal{U}$ are defined as subsets of the Cartesian product $\mathcal{U}^k$. Relations play an important role in classification problem. For example, the distance-based methods like nearest neighbor classifiers or clustering are based mainly on similarity relation between objects defined by the distance function.

Investigations on concept approximation problem are well motivated both from theoretical as well as practical point of view [7] [8]. As an example, let as remind that the standard rough sets were defined by indiscernibility between objects which is an equivalence relation, while similarity relation approximating the indiscernibility relation is the tool for many generalizations of rough set theory including the tolerance approximation space [9], similarity based rough sets [10], rough set methods for incomplete data [11], rough set methods to preference-ordered data [12] [13].

In this paper we investigate the problem of searching for approximation of relations from data. We show that this method is the basic component of many compound tasks. We also present a novel rough set based approach to discovering useful patterns from nonstandard and complex data for which the standard inductive learning methodology fails. The proposed solution is based on a two-layered learning algorithm. The first layer consists of methods that are responsible for searching for (rough) approximation of some relations between objects from the data. At the second layer, the approximated relations induced by the first layer are used to synthesize the solution of the original problem. The critical problem in any layered learning system is how to control the global accuracy by tuning the quality of its components. We present a solution of this problem based on the changing of the quality of approximate relations.

We describe two representative examples related to binary relations to demonstrate the power of the proposed methodology. In the first example, we consider the problem of extracting the optimal similarity relation and present some applications of approximate similarity relations in classification problem. We present the advantages of this method comparing with the standard classification methods [14] [15].

The second example relates to the approximation of preference relation and its applications in (1) learning ranking order on a collection of combinations, (2) predicting the values of continuous decision attribute, (3) optimizing the process of searching for the combination with maximal decision [16]. This method can be applied to mining ill-defined data, i.e., data sets with few objects but a large number of attributes. Results of some initial experiments on medical and biomedical data sets were very promising.

**Keywords:** Rough sets, relation approximation, knowledge discovery.

## Acknowledgements

This research was partially supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

## References

1. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11** (1982) 341–356
2. Pawlak, Z.: Some issues on rough sets. Transaction on Rough Sets **1** (2004) 1–58
3. Bazan, J., Nguyen, H.S., Szczuka, M.: A view on rough set concept approximations. Fundamenta Informatica **59**(2-3) (2004) 107–118
4. Bazan, J.G., Nguyen, S.H., Nguyen, H.S., Skowron, A.: Rough set methods in approximation of hierarchical concepts. In Tsumoto, S., Slowinski, R., Komorowski, H.J., Grzymala-Busse, J.W., eds.: Rough Sets and Current Trends in Computing: Proceedings of RSCTC'04, June 1-5, 2004, Uppsala, Sweden. Volume LNAI 3066 of Lecture Notes in Computer Science., Springer (2004) 346–355
5. Nguyen, S.H., Bazan, J., Skowron, A., Nguyen, H.S.: Layered learning for concept synthesis. In Peters, J.F., Skowron, A., Grzymala-Busse, J.W., Kostek, B., Swiniarski, R.W., Szczuka, M.S., eds.: Transactions on Rough Sets I. Volume LNCS 3100 of Lecture Notes on Computer Science. Springer (2004) 187–208
6. Stone, P.: Layered Learning in Multi-Agent Systems: A Winning Approach to Robotic Soccer. The MIT Press, Cambridge, MA (2000)
7. Skowron, A., Pawlak, Z., Komorowski, J., Polkowski, L.: A rough set perspective on data and knowledge. In Kloesgen, W., Żytkow, J., eds.: Handbook of KDD. Oxford University Press, Oxford (2002) 134–149
8. Stepaniuk, J.: Optimizations of rough set model. Fundamenta Informaticae **36**(2-3) (1998) 265–283
9. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**(2-3) (1996) 245–253
10. Slowinski, R., Vanderpooten, D.: Similarity relation as a basis for rough approximations. In P., W., ed.: Advances in Machine Intelligence & Soft-computing, Bookwrights, Raleigh (1997) 17–33
11. Greco, S., Matarazzo, B., Słowiński, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In Zanakis, S., Doukidis, G., Zopounidis, C., eds.: Decision Making: Recent Developments and Worldwide Applications. Kluwer Academic Publishers, Boston, MA (2000) 295–316

12. Slowinski, R., Greco, S., Matarazzo, B.: Rough set analysis of preference-ordered data. In Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N., eds.: Third International Conference on Rough Sets and Current Trends in Computing RSCTC. Volume 2475 of Lecture Notes in Computer Science., Malvern, PA, Springer (2002) 44–59
13. Slowinski, R., Greco, S.: Inducing robust decision rules from rough approximations of a preference relation. In: ICAISC. (2004) 118–132
14. Nguyen, S.H.: Regularity analysis and its applications in data mining. In Polkowski, L., Lin, T.Y., Tsumoto, S., eds.: Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Volume 56 of Studies in Fuzziness and Soft Computing. Springer, Heidelberg, Germany (2000) 289–378
15. Wojna, A.: Analogy based reasoning in classifier construction. (In: Transactions on Rough Sets IV: Journal Subline) 277–374
16. Nguyen, H.S., Łuksza, M., Mkosa, E., Komorowski, J.: An Approach to Mining Data with Continuous Decision Values. In Klopotek, M.A., Wierzchon, S.T., Trojanowski, K., eds.: Proceedings of the International IIS: IIPWM05 Conference held in Gdansk, Poland, June 13-16, 2005. Advances in Soft Computing, Springer (2005) 653–662